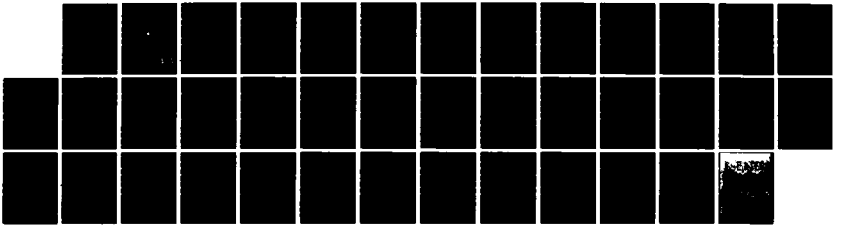
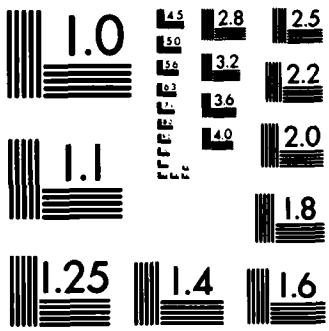
AD-A130 252 RELIABILITY OF SLOPE SCORES FOR INDIVIDUALS(U) NAVAL BIODYNAMICS LAB NEW ORLEANS LA R C CARTER ET AL. APR 83 NBDL-83R003 1/1

UNCLASSIFIED F/G 5/10 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

NBDL - 83R003

Reliability of Slope Scores for Individuals

Robert C. Carter and Michele Krause

April 1983

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NBDL-83R003	2. GOVT ACCESSION NO. AD-A130252	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Reliability of Slope Scores for Individuals		5. TYPE OF REPORT & PERIOD COVERED Research Report
		6. PERFORMING ORG. REPORT NUMBER NBDL-83R003
7. AUTHOR(s) Robert C. Carter and Michele Krause		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Biodynamics Laboratory Box 29407 New Orleans, LA 70189		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS MF58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research & Development Cmd National Naval Medical Center Bethesda, MD 20014		12. REPORT DATE April 1983 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

slope scores, reliability, repeated measures, human information processing

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Slope scores representing the rate of human information processing have often been used as dependent variables in experiments and in correlation studies. The reliability of the slope scores for individuals is an important consideration because it affects the power of experiments and the maximum expected correlation in correlation studies. This article examines the reliabilities (inter-day correlations) across 15 days of repeated measurements for each of six prominent human information processing tasks: high speed memory scanning, proactive memory interference, semantic reasoning, letter search, typographic error

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

Block 20 (con'td)

search, and choice reaction time. In each case, the reliability of the slope scores is less than the reliabilities of the mean response times from which the slopes were calculated. This is remarkable because the slopes include more data than each mean response time. Reasons for the relative unreliability of slope scores are discussed. Strategies for improving the reliability of slope estimates are suggested. Finally, it is argued that in applied experimental research it is usually unnecessary to calculate slope scores for individuals because the more reliable mean response times are sufficient to answer common applied questions.

NBDL - 83R003

Reliability of Slope Scores for Individuals

Robert C. Carter and Michele Krause

April 1983

Bureau of Medicine and Surgery
Work Unit No. MF58.524-002-5027

Approved by

Channing L. Ewing, M. D.
Chief Scientist

Released by

Captain L. E. Williams, MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

SUMMARY PAGE

PROBLEM

It is sometimes necessary to compare the human performance effects of alternative equipment designs or work environments, or to evaluate the performance effects of training or occupational therapy. When these comparisons and evaluations are made with repeated measurements, the capability of the experiment to show an effect of the alernatives or the training depends partly upon the reliability of the human performance criterion being measured. Recently, least-squares regression slopes derived from "information-processing" tasks have come into vogue as performance measures. Their utility in applied experiments depends upon their reliability, which has been suggested to be relatively low.

FINDINGS

Experimental evidence, based on repeated measures of prominent human information-processing tasks, indicated that slope scores are consistently less reliable than mean response time scores.

RECOMMENDATIONS

Mean response time should be chosen over slope scores in applied repeated measures experimentation.

This research work was funded by the Naval Medical Research and Development Command.

The volunteers used in this study were recruited, evaluated and employed in accordance with the procedures specified in the Secretary of the Navy Instruction 3900.39 series and the Bureau of Medicine and Surgery Instruction 3900.6 series. These instructions are based upon voluntary consent, and meet or exceed the prevailing national and international guidelines.

Abstract

Slope scores representing the rate of human information processing have often been used as dependent variables in experiments and in correlation studies. The reliability of the slope scores for individuals is an important consideration because it affects the power of experiments and the maximum expected correlation in correlation studies. This article examines the reliabilities (inter-day correlations) across 15 days of repeated measurements for each of six prominent human information processing tasks: high speed memory scanning, proactive memory interference, semantic reasoning, letter search, typographic error search, and choice reaction time. In each case, the reliability of the slope scores is less than the reliabilities of the mean response times from which the slopes were calculated. This is remarkable because the slopes include more data than each mean response time. Reasons for the relative unreliability of slope scores are discussed. Strategies for improving the reliability of slope estimates are suggested. Finally, it is argued that it is usually unnecessary to calculate slope scores for individuals because the more reliable mean response times are sufficient to answer common research questions.

Reliability of Slope Scores for Individuals

Human information processing is an analogy which likens mental events to data processing by a computer. The general method derived from this analogy is to measure the time required for various numbers of iterations of a presumed mental process, and to calculate the time per iteration (Posner, 1978). The time per iteration is commonly expressed as a slope score. Although the method has been widely and successfully employed by experimental psychologists, one question about the human information processing analogy has been underemphasized. The time required by a computer to complete a particular computation is highly consistent from one day to the next. How reliable is the rate of mental events? What implications does the reliability of the rate have for human information processing research?

The purpose of this article is to show whether the slope score for individuals is a reliable indicator of human information processing. Memory scanning (Sternberg, 1966), interference susceptibility (Underwood, Boruch, & Malmi, 1977), semantic reasoning (Collins & Quillian, 1969), letter search (Neisser, 1963), search for typographical errors in prose (Schindler, 1978), and choice reaction time (Teichner, 1978) were studied with respect to the reliability of the slope score in repeated measures experiments. Implications of the results for measurement of individual differences, and repeated-measures experiments on effects of interventions (e.g., changes of training, equipment design, or work environment) will be discussed.

Reliability is the extent to which a person's score on a given occasion of performance testing is predictable from a previous performance. Recall Thorndike's (1947) classification of types of variance in a person's behavior. The classes of variance were: (a) lasting and general (i.e., ability), (b) lasting and specific (e.g., knowledge of a particular item on one test form), (c) temporary but general (e.g., general state of motivation), (d) temporary and specific (e.g., attitude toward a particular type of test item), and (e) chance (e.g., guessing). A measurement of human performance is reliable to the extent that class (a) variance is large relative to all other classes. Unreliable performance measurements represent mere transients, and hence are of limited interest and usefulness; they are unrelated to behavior in the future.

The extent to which performance is reliable, relative to other people's performances, is represented by intertrial correlations of performance scores. Intertrial correlations are product-moment correlations between scores in each pair of repeated measurements. They represent the extent to which subjects maintain the same positions (ranks) relative to each other from one trial to another. They can be increased by reducing errors of measurement and by increasing the range of test skill represented by the subjects. The power of repeated measures analysis of variance increases with reliability (Sutcliffe, 1980).

Some will say that arguments regarding slope scores are irrelevant if they rest on considerations of reliability. After all, the choice of a slope score is based on theoretical, not reliability, considerations. The regression slope coefficients have meaning in the context of information-

processing models. (Similar reasons were given years ago for using difference scores, according to Cronbach and Furby (1970)). The usefulness of the slope concept is not questioned here. Rather, it is pointed out that there are alternative methods for implementing the slope concept, each having a different expected reliability. One possible method which is commonly used is to calculate slope scores for each individual, and then to use the slopes as a dependent variable or as a variable in a correlation study. It is hoped that the reader will become convinced that this will result in less powerful experiments and correlation studies than the procedure of simply using response times. The recommended alternative method is that slopes be calculated using mean response times for the group of subjects rather than for each subject individually. The slope score for each experimental condition (e.g., age groups, equipment designs, or environmental stressors) would be the linear contrast of mean response times in that condition. Response time, instead of the slope score for each individual, is to be preferred as the dependent variable because response time is more reliable than the slope score calculated from it. This remarkable fact is not apparent a priori. Furthermore, these two alternative methods yield equal estimates of the slope. The mean of individual subjects' slope scores will equal the slope of their mean response times. For purposes of comparing slopes associated with various experimental treatments, or for correlation studies, slopes of mean response times are preferable to slope scores for individuals on the basis of reliability.

Reasons for Unreliablity of Individual's Slopes

There are several reasons why slope scores might be expected to be unreliable. Briefly, the reasons are incorrectness of the linear model, restriction of range of slope scores, sensitivity of the slope score to outliers, and the kinship between slope scores and difference scores. Not all of these reasons will be operative in all applications of slope scores, and some of them are interrelated. Each of the reasons will be considered in the following paragraphs.

Incorrectness of the Linear Model

In some cases in which slope scores have been reported, there is at least a little doubt that the relation between response time (RT) and the independent variable is linear. To the extent that the linear model is incorrect, there is a component of the variance of the slope which is error variance associated with model bias (Draper & Smith, 1966). Recall that the correlations which represent reliability are the ratio of true score variance divided by true score plus error variance. As the error variance increases due to model bias, the reliability of the slope score will approach zero. Of course this will not be a problem if the linear model is accurate.

Restriction of Range of Slope Scores

It may occasionally happen that subjects' response-time scores retain the same relative positions at each level of the independent variable. This condition is indicated by a correlation between the RTs at extremes of the independent variable which approaches unity when corrected for attenuation. Stated more graphically, the plots of subjects' response times versus the independent variable are practically parallel. This means that the sub-

slopes are all very nearly equal. Such a finding suggests that the slope under consideration is invariant across individuals, so no reliable individual differences of slopes exist.

Sensitivity of the Slope Score to Outliers

Least squares slopes are very sensitive to outliers at the extremes of the independent variable (Draper & Smith, 1966). This problem has received attention in discussions of the statistical robustness of the product-moment correlation (e.g., Devlin, Gnanadesikan, & Kettenring, 1975). These discussions are relevant to slopes because the slope is simply the correlation multiplied by the ratio of the standard deviation of the dependent variable divided by the standard deviation of the independent variable. The conclusion one draws upon reading about the lack of robustness of the correlation coefficient is that a slope score calculated for an individual should be regarded with caution. Hence, a set of slope scores from several subjects on one occasion is likely to include an erroneous slope, and consequently, it would be surprising to see a successful duplication of several subjects' slope scores on two or more occasions of measurement.

Kinship Between Slope Scores and Difference Scores

It is well known that difference scores tend to be unreliable (c.f., Cronbach & Furby, 1970). The concept of slope is defined algebraically as the ratio of the difference in the dependent variable divided by the difference of the independent variable. This algebraic definition of slope is identical with the least-squares procedure for calculating slopes if data are taken at two levels or three equally-spaced levels of the independent variable.

The claim that slopes and differences are related will be demonstrated algebraically for data obtained at three equally-spaced levels of the independent variable ($\underline{X}$). The least-squares slope is calculated as $\underline{XY}/\ \underline{X}^2$ if $\underline{X}$ and $\underline{Y}$ (the dependent variable, e.g., RT) are scaled to have zero means. This slope formula expands to:

$$(\underline{X}_1\underline{Y}_1 + \underline{X}_3\underline{Y}_3)/(\underline{X}_1^2 + \underline{X}_3^2).$$

But $\underline{X}_1 = -\underline{X}_3$. Hence, the slope is:

$$\underline{X}_1(\underline{Y}_1-\underline{Y}_3)/2\underline{X}_1^2 \text{ or } (\underline{Y}_1-\underline{Y}_3)/2\underline{X}_1.$$

In addition, without any loss of generality, $\underline{X}$ can be scaled so that its extreme values ($\underline{X}_1$ and $\underline{X}_3$) are plus and minus one-half. Therefore, the least-squares slope is a difference score ($\underline{Y}_1-\underline{Y}_3$) in the case of data obtained at three equally-spaced levels of the independent variable, a common case. Furthermore, it can be shown that the reliability of $\underline{Y}_1-\underline{Y}_3$ is less than the reliability of $\underline{Y}_1$ or $\underline{Y}_3$ under ordinary conditions (Lord & Novick, 1968). Using this fact and the algebraic derivation we have the inference that, at least for this case, the reliability of the least-squares slope is less than the reliability of its component response times.

The identical relationship between slopes and differences for data at two or three equally-spaced levels fades to a more remote kinship as the number of levels of the independent variable increases. Nonetheless, the identity of slope and difference scores in the two and three-level cases indicates that slope scores suffer from reliability problems similar in origin to those of difference scores.

Next we will illustrate experimentally the relative unreliability of slope scores for individuals, compared with response times from which the slopes were calculated.

SIX INFORMATION PROCESSING EXPERIMENTS

Methods Common To All Six Experiments

Subjects

The subject pool was common to all experiments. The subjects in these experiments were males enlisted in the U.S. Navy. They were carefully selected to be in good health. Details of the subject selection procedures were given by Thomas, Majewski, Ewing, and Gilbert (1978). The subjects had volunteered for participation in accordance with federal and international guidelines on informed consent.

Apparatus and Procedure

The experimental stimuli in experiments 1, 2, and 3 were presented on a Kodak Audioviewer® which is a self-contained slide projector and rear-projection screen (22 cm square). Subjects viewed the screen from a distance of about 50 cm. The stimuli (slides) were timed by markers on a cassette tape read by the control unit of the Audioviewer®. The experimenters programmed the tape according to specifications given in the description of each experiment. The subjects' responses were timed to the nearest 5 msec by an electronic timer which was started by presentation of a stimulus item and stopped by a button-push response. The subjects performed each information-processing task once each day for three weeks (Saturdays and Sundays excluded). Hence the data for each task included scores for each subject on each of 15 successive weekdays.

Experiment 1: High Speed Memory Scanning

Method

The high speed memory scanning task was administered to 23 subjects in general conformity with Sternberg's (1966, 1969, 1975) descriptions. Subjects were presented with a slide containing 1, 2, 3, or 4 digits (the positive set) each subtending about 1.5 degrees of visual angle. (All digits not presented constituted the negative set.) The duration of this stimulus was one second per digit. A probe digit followed presentation of the positive set by two seconds. The subject was to select one of two responses, depending upon whether the probe was from the positive or negative set. Response time (RT) was recorded.

Each daily session of testing included ten responses for each positive set size. Trials for each set size were in blocks. The order of the blocks was the same each day, ascending from one- to four-item positive sets. Each block included five probes from the positive set, and five from the negative set. The digits of the positive set and the probe digits were chosen at random for each trial, but were the same for all subjects.

The RT were used to calculate slope and intercept scores for each subject on each day, in accordance with Sternberg's (1966, 1969) finding that RT increases linearly with positive set size. According to Sternberg (1966, 1969, 1975) the slope may be interpreted as the rate of search through short-term memory, and the intercept represents the time required for stimulus processing and response formulation.

Slope and intercept scores were computed using least-squares regression. There was a regression equation for each subject on each day which expressed the 40 RT for that subject on that day as a linear function of positive set size. Daily means, standard deviations, and interday correlation matrices (all calculated across subjects) were developed for each of the following scores: subject mean RT for positive set sizes 1, 2, 3, and 4; slope of mean RT versus positive set size; intercept; and percent error.

Results

Slopes and intercepts. In this experiment, the intercept (450 msec) did not change appreciably over 15 days ($F(14,280) = 1.53$, $p > .1$) and is comparable to that (397 msec) reported by Sternberg (1966). The variance of the intercept, some of which is attributable to individual differences ($F(14,280) = 14.25$, $p < .005$), did not change with practice ($F_{max}(15,20) = 4.08$, $p > .05$) (David, 1952). However, the slope scores obtained in this experiment decreased with practice ($F(14,280) = 5.32$, $p < .005$). This is a common finding (Kristofferson, 1972; Ross, 1970; Simpson, 1972; Burrows & Murdock, 1969). The mean slopes did not change appreciably after the second day of testing ($F(12,280) = 1.33$, $p > .25$). The average slope obtained after the second day in this experiment (44.2 msec/item) is similar to the average slope obtained by Sternberg (1966) with practiced subjects (37.9 msec/item). In contrast to the mean of the slopes, which changed with practice, the variance of the slopes remained about the same from day to day. ($F_{max}(15,20) = 3.74$, which does not exceed the critical point ($F_{max}(15,20) = 4.75$, $p = .05$) calculated according to David (1952). Apparently, an appreciable part of the slope variance within each day was attributable to differences among the subjects ($F(20,280) = 2.57$, $p < .005$), rather than measurement errors. Table 1 displays the intertrial (interday) correlation matrix for the slope score, $r = .11$. It indicates that a person's slope score on one day is virtually useless for prediction of performance on another day, relative to other subjects. The implications of this unreliability for measurement of individual differences and repeated measures experiments will be discussed later.

Table 1
Item Recognition: Slope Reliabilities (x 100) over 15 days (n=21)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-13	03	-09	16	21	-21	28	42	21	36	-22	21	15	24
2		02	-10	-03	44	01	41	-38	-07	19	61	31	05	21
3			-06	31	19	-17	-07	26	02	45	03	11	-43	11
4				39	23	54	01	12	-07	28	-20	05	-02	-02
5					31	02	-11	47	37	65	-32	37	04	-08
6						-37	34	-02	03	43	36	53	25	12
7							-07	-14	-05	04	-09	-24	-04	-07
8								21	-19	03	40	58	21	28
9									40	37	-56	25	09	03
10										30	-57	-02	19	01
11											-19	33	-03	28
12												42	-05	01
13													14	17
14														-09

Error rates. The present results contrast with Sternberg's (1966) in that the subjects' error rate was 6%, rather than the 1.3% he obtained. The mean error rate was unaffected by practice ($F(14,280) = .8$, p .3). An important point for interpretation of the slope and RT scores is that the mean error rate was independent of positive set size ($F(3,60) = .16$, p .5). The variance of the error rate was unchanged over the 15 days of the experiment ($F\max(15,20) = 2.90$, p .05) and was significantly attributable to individual differences ($F(20,280) = 168.08$, p .005).

Response times. Mean RT across subjects and days for each positive set size is presented in Table 2. Neither the mean nor the variance of RT for positive set size one (RT1) was affected by practice (F (14,280) = 1.20, p = .28 and Fmax(15,20) = 2.65, p .05, respectively). Similarly, the mean (across all days) and variance (during the first 14 days of the experiment) of RT2 were unaffected by practice (F(14,280) = .97, p = .49 and Fmax (14,20) = 3.05, p .05, respectively). Unfortunately, one subject responded extraordinarily slowly to positive-set-size-two on Day 15, so that the variance on that day was triple the magnitude of the next largest daily variance. The means of RT3 and RT4 were affected significantly by practice (F(14,280) = 4.19, p .005 and F(14,280) = 13.14, p .005, respectively). This effect of practice on RT3 and RT4 became undetectable after Day 2 (F(12,280) = 1.02, p .05 and F(12,280) = 1.30, p .05, respectively). The variance of neither RT3 nor RT4 were significantly affected by practice (Fmax(15,20) = 4.70, p .05 and Fmax(15,20) = 4.65, p .05, respectively), although the variances for the first two days were the largest in each case. An aspect of Table 2 which is worth noting is that the interval between RT1 and RT2 is greater than that between RT2 and RT3 or RT3 and RT4. If the linear (slope) model advocated by Sternberg (1966, 1969, 1975) were true, then RT2-RT1 = RT3-RT2 = RT4-RT3. Clearly this is not the case ($F_{nonlinear}$ (2,40) = 8.59, p .01), so a slope may not be an appropriate way to represent these data.

Table 2: Means and Standard Deviations Obtained in a Memory Scanning Experiment (Experiment 1)

Size of Positive Set			
1	2	3	4
588(28)	678(31)	692(28)	724(29)

Note: Response time in milliseconds; standard deviations in parentheses. Data averaged over the last 13 days of a 15-day experiment.

Table 3 discloses the intertrial correlation for RT1 (above the diagonal) and RT4 (below the diagonal), $\bar{r}$ = .69 and .72, respectively. The intertrial correlation matrices for RT2 and RT3 were similar. Note that the intertrial correlations for RT are much larger and more homogeneous than those for the slopes derived from RT. Hence, the poor task definition and instability of the slope scores (indicated by the intertrial correlation matrix of the slopes, Table 1) cannot be blamed on unreliability of the RT from which the slopes were calculated.

Table 3: Memory Scanning Intertrial Correlations (X 100): Positive Set 1 Above Diagonal, Positive Set 4 Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		74	64	64	59	69	65	51	38	44	51	65	59	59	64
2	48		86	63	85	88	80	82	64	56	61	86	74	74	71
3	56	61		62	91	77	66	65	54	46	63	74	65	70	67
4	60	53	78		74	77	78	57	53	46	57	70	62	56	52
5	57	41	82	85		84	75	75	67	53	71	79	67	72	56
6	57	45	73	78	84		91	84	79	66	70	96	85	82	79
7	59	39	73	77	84	83		71	73	52	53	88	75	69	75
8	67	39	72	75	87	80	79		84	80	73	83	76	76	60
9	44	25	64	69	87	87	82	85		70	62	73	68	64	64
10	50	45	76	80	91	80	78	90	88		61	63	70	52	51
11	57	48	84	91	85	71	75	76	71	83		63	74	84	53
12	54	40	82	81	92	83	77	86	86	88	90		81	81	76
13	60	26	63	71	86	81	80	86	91	88	74	83		72	72
14	59	46	66	66	87	78	77	90	88	91	69	81	92		68
15	56	20	65	76	88	68	81	80	81	81	86	86	85	78	

One further question might be asked regarding the RT scores. Does the level of performance with one positive set size tell us anything about a person's level of performance with another positive set size? If performance for RT4, say, were perfectly correlated with performance for another RT, then the other RT would be superfluous because it is predictable from the RT4 measurement. For example, the average correlation between RT4 and RT1 ($\bar{r}$ = .74) was large. Indeed, when corrected for attenuation ($\bar{r}$ = .96) it was nearly perfect. Similar, but more dramatic results were obtained with the other 5 pairs of RT for positive set sizes that are less disparate than 1 and 4. Almost all (94%) of the reliable variance of response time to positive set size one was predictable from response time to positive-set-size-four. This fact will later help to explain why the slope score was so unreliable, and hence undesirable as a dependent variable in repeated measures experiments and applications of individual differences.

Summary of results for high speed memory scanning. Results of analyses of the means, variances, and intertrial correlations were reported for slope, intercept, percent errors, and RT to positive set sizes 1, 2, 3, and 4. The intertrial correlations of the slope score were near zero, indicating no temporal generalizability of that score. However, the RT from which the slopes were calculated possessed encouraging reliability. Finally, the response times to various positive set sizes were almost as well correlated with each other as they possibly could be. Therefore, RT1, 2, 3, and 4 measure the same personal attribute, so three of the set sizes were superfluous unless the 6% reliable variance which was unique to a particular RT is of interest.

Experiment 2: Interference Susceptibility

The second task selected for study was Underwood's Interference Susceptability Test (Underwood et al., 1977). This task was originally designed by Underwood et al. to study the effects of proactive interference in memory. In the original study, 200 college students were tested on 24 separate tasks. A slope score was planned to represent the increase of proactive memory interference with repeated exposure to the same memory items, but the slope score was found to be unreliable. Fernandes and Rose (1978) also included the test in their studies of an information-processing approach to performance assessment.

Method

Twenty-three subjects participated in this experiment. Stimulus material was comprised of lists of trigram-digit pairs (e.g. NOB-2). A list was made up of five trigrams paired with digits from 1 to 5. During each session, three sets, each containing four lists, were administered. Across the four lists of a set, the same trigrams were paired with digits from 1 to 5, forming different combinations in each list.

Subjects were shown each of five trigram-digit pairs by means of a single slide. The rate of presentation was one slide every 3 seconds. A cueing slide appeared at the end of the list and at the beginning of the recall list. Each trigram was then shown by itself for 4 seconds (in an order different from the paired presentation) and subjects recorded the number with which they thought each trigram had been paired.

Results

Two measures were taken across sets for four lists: (a) slope of lists and (b) percent correct for each list. Table 4 shows the mean percent correct responses across sets for the four lists. As expected, performance declined with each successive list (within a set) that was presented. The average percent correct in both this study and Fernandes and Rose (1978) study was 65%. Underwood, et al. (1977) obtained an 85 percent correct average when they studied this test and 23 others in a battery of memory tests.

Table 4: Means and Standard Deviations Obtained in a Proactive Interference Susceptibility Experiment (Experiment 2)

Order of Lists			
1	2	3	4
76.46(8.30)	69.50(6.25)	61.63(6.00)	60.23(6.72)

Note: Percent correct; standard deivations in parentheses. Data averaged over 3 sets per list for the last 13 days of a 15-day experiment.

When Underwood et al. (1977) correlated total correct responses for sets 1, 3, 5 with those from sets 2, 4, 6, they obtained a value of $r = .81$. This correlation between successive sets (i.e., split half) in Underwood's study is compared to a correlation of $r = .74$ between successive days (i.e., test-retest) in the present research, wherein the number of observations are the same for both calculations. There is no evidence that the reliabilities of the present data are different from those of Underwood et al. (1977) ($z = .72$, p .40). Table 5 shows reliabilities within Lists 1 and 4, $r = .46$ and .32, respectively. Lists 2 and 3 gave comparable results. Table 6 shows reliabilities for the mean list slope across sets. Composite reliability for this score is essentially zero ($r = .09$). There were significant differences among the subject's slopes ($F(2, 308) = 2.45$, p .005), so the low reliabilities were not due to restriction of range.

To summarize, the chief finding in this experiment is that the slope score, theoretically the most meaningful measure of the interference factor, is unreliable. Fernandes and Rose (1978) also obtained low reliability for the slope measure ($r = .05$). However, mean number of correct responses were moderately reliable.

Table 5: Interference Susceptibility Intertrial Correlations (X 100): List 1 Above Main Diagonal, List 4 Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		60	71	53	33	16	29	30	33	31	19	52	99	22	36
2	32		72	54	47	12	42	36	38	50	39	52	17	40	38
3	59	05		67	51	39	39	30	41	54	50	59	28	43	52
4	40	10	34		66	43	66	10	56	36	64	80	59	35	57
5	38	-07	18	33		-00	59	27	38	43	73	77	45	36	50
6	-16	18	-43	-01	23		39	09	54	24	28	33	55	43	22
7	34	43	14	40	40	05		45	62	42	57	72	52	63	47
8	08	08	-05	23	48	00	55		37	73	30	38	16	48	21
9	60	15	16	47	26	15	35	18		60	59	62	68	71	48
10	33	52	05	25	48	06	59	34	40		56	55	41	70	43
11	32	42	-06	30	35	22	81	49	34	44		78	62	58	61
12	58	35	34	15	24	-15	47	09	45	61	40		56	59	58
13	39	16	04	42	61	16	42	37	50	60	55	40		59	60
14	48	14	06	51	55	10	52	55	64	53	54	41	73		60
15	44	20	10	50	50	27	46	02	37	49	46	41	59	44	

Table 6: Interference Susceptibility Intertrial Correlations (X 100) of Mean List Slope Across Sets

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-21	15	-14	04	03	04	-17	16	06	-08	08	14	-15	-04
2		10	-02	-13	-05	32	-09	-05	10	29	09	08	-05	02
3			07	16	-10	-05	30	-01	32	-31	16	01	23	-28
4				-19	20	39	-15	-11	-29	01	-06	25	19	03
5					-31	-13	34	06	42	32	40	57	56	06
6						32	-12	23	-29	-08	-14	07	-09	04
7							-03	28	-08	54	40	38	03	18
8								-08	24	09	-18	-15	26	-15
9									16	08	49	29	-01	00
10										01	35	17	14	33
11											45	42	-11	29
12												60	29	14
13													39	07
14														03

Experiment 3: Semantic Reasoning

Method

A semantic reasoning task was administered to 23 subjects in the manner of Collins and Quillian (1969). Subjects were presented with a sentence (via photographic slide) describing either a property relation (P) (e.g., "pepper is hot") or a superset relation (S) (e.g., "an apple is a fruit".) Each P or S sentence was labeled as either a 0, 1, or 2, denoting the suspected amount of searching through memory involved in making a decision. For example, the sentence "an apple is an apple" is a S0 sentence because it does not require any indexing through memory to determine whether this sentence is true or false. An S1 sentence (e.g., "an apple is a fruit") is thought to require one "step" through memory and a S2 (e.g., "an apple is a food") is suspected to require even more memory searching. Property sentences (P0, P1, P2) were also presented.

Four sentences of each type (P0, P1, P2, S0, S1, S2) were given for a total of 24 items during each testing session on each of 15 successive weekdays. Half of the sentences were true and half false. The stimulus items for each session were chosed at random from a pool of 144 items. Each sentence remained on the screen for 4 seconds. Response times were recorded. The intertrial interval was approximately 3 seconds.

Table 7: Means and Standard Deviations Obtained in a Semantic Reasoning Experiment (Experiment 3)

Memory level:	0	1	2
Superset:	1.15(.16)	1.26(.20)	1.31(.25)
Property:	1.35(.16)	1.39(.14)	1.41(.18)

Note: Response time in seconds; standard deviations in parentheses. Data averaged over the last 14 days of a 15-day experiment.

Results

Table 7 shows the mean RT for P and S sentences of types 0, 1, and 2. As noted by Collins and Quillian (1969), RT increased with the number (0, 1, or 2) of hypothetical steps through memory needed to verify the sentence. This was true for both P and S sentences. Collins and Quillian suggested that this increase be represented by a slope.

Table 8 lists intertrial correlations for Property 0 sentences (above the main diagonal, $\underline{r}$ = .46) and for Superset 0 sentences (below the main

diagonal, $\bar{r}$ = .65). Similar results were obtained for type 1 and 2 sentences. These correlations, although weak, are much healthier than the correlations in Table 9 of slopes across sentence types 0, 1, and 2 for P (above the diagonal, $\bar{r}$ = .01) or S (below the diagonal, $\bar{r}$ = .02) sentences. In this case, the unreliability of the slope score may have been due to restriction of range. The differences in slope scores among subjects were small compared with their errors of measurement: for P sentences ($F(20,280) = 1.02$, $p = .44$); and for S sentences ($F(20,280) = .91$, $p = .58$). In any case the slope scores for memory steps were unreliable compared with the response times from which they were generated.

Table 8: Semantic Reasoning Intertrial Correlations (X 100) of Response Time: Property 0 Above Diagonal, Superset 0 Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		52	53	23	77	57	48	64	28	59	62	44	35	29	24
2	68		60	29	46	53	58	70	57	16	75	50	55	09	30
3	65	66		33	58	45	61	66	60	-05	69	61	62	51	69
4	54	58	83		28	42	49	43	19	29	14	53	43	20	26
5	57	56	76	78		81	73	65	30	62	74	55	58	53	25
6	74	68	69	69	75		79	72	28	62	58	56	72	43	14
7	83	62	67	67	72	78		72	41	36	61	43	59	50	29
8	66	49	69	69	69	84	70		35	39	72	67	69	53	56
9	69	52	80	80	63	59	63	54		-24	49	26	50	17	40
10	81	70	62	62	59	86	84	74	62		26	24	18	18	-32
11	61	49	68	68	62	80	61	75	52	72		58	68	40	42
12	66	62	50	50	48	77	69	76	52	81	59		75	30	53
13	77	43	47	47	54	46	75	56	68	49	47	55		39	48
14	72	62	44	44	55	73	78	61	46	85	75	68	52		44
15	60	58	70	70	75	71	72	61	54	75	79	59	46	76	

Table 9

Semantic Reasoning Slope Reliabilities (X 100):
Property Above Diagonal, Superset Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		-24	-13	23	45	-02	-06	-01	11	-13	-15	-22	-15	00	-09
2	08		12	-23	-70	06	-15	-53	-44	-25	40	-38	-07	-20	09
3	-38	27		-27	-03	-08	-23	-04	10	07	21	-08	06	-14	62
4	-35	-01	21		-08	22	61	00	31	08	12	-14	14	-22	-03
5	-14	38	-07	23		-07	-16	35	10	15	-34	18	-01	21	-15
6	20	03	-22	-60	06		20	23	11	41	-08	-11	24	16	01
7	41	40	-08	08	02	23		-13	27	24	16	08	-04	14	-05
8	14	-34	-28	-24	-52	-18	-35		18	51	-59	64	51	-13	13
9	-03	39	26	21	52	-01	-04	-54		-30	-38	19	-13	-22	23
10	16	49	02	-46	17	47	53	-38	24		13	23	21	25	10
11	16	00	19	-68	-13	34	-18	05	01	24		-50	-23	12	-01
12	44	11	-16	-02	-05	22	49	10	-17	11	-10		32	-01	-03
13	29	27	-06	52	19	-40	34	-16	10	-14	-33	-15		-34	20
14	-11	-40	-02	24	-42	-30	-53	45	-19	-76	-24	-18	11		-49
15	-10	53	30	22	22	09	37	-25	-05	-08	-02	06	37	-14	

Experiment 4: Letter Search

Method

A letter-search task was administered to 23 subjects in the manner of Neisser (1963). Subjects were presented with four columns of 16 groups of five upper case letters typed on an 8½ x 11-inch sheet of paper. They were to scan the columns and put a check mark next to any group that contained a pre-announced target letter. Each group included a target letter with probability .50.

The number of possible target letters announced was 1 or 2 or 4 in the three parts of this task. Each part was done on a separate sheet of paper. Subjects were allowed 20 seconds to work on the 1-target sheet, and 30 seconds to work on each of the 2 and 4-target sheets. The test was repeated in 1, 2, and 4-target order on each of 15 successive weekdays, and two test forms were used on alternate days. Neisser, Novick, and Lazar (1963) found that search time increased linearly with the number of possible targets. They used a least squares regression slope to represent the rate at which decision making time increased as the number of alternatives increased.

Results

Table 10 shows the mean response times per group of five letters for the one, two, and four-target parts of this task. In the aggregate, the linear model proposed by Neisser appears to be excellent. Response Time (seconds) = .195 + .335 (The Number of Targets) is the model fit to our data, and it explains 99.9 percent of the variance of the means listed in Table 10.

Table 10: Means and Standard Deviations Obtained in a Letter Search Experiment (Experiment 4)

Number of Targets		
1	2	4
.54(.05)	.85(.06)	1.54(.14)

Note: Response time in seconds per item; standard deviations in parentheses. Data averaged over the last 13 days of a 15-day experiment.

Table 11 displays the intertrial correlations of the response times for one-target (above the main diagonal) and four-target (below the main diagonal) tasks. The average one-target correlation is .58, and the average four-target correlation is .44. The intertrial correlations for the slope scores are shown in Table 12. Once again, the slope reliability is considerably lower than that of the response times of which it is composed. The average slope intertrial correlation is .30. In the case of this letter-search task, the linear model appears to fit the data very well, and there are appreciable differences among subjects' slope scores ($F(14,308) = 6.47$, $P < .005$), so model bias or restriction of range are unlikely explanations of the slope score's relatively poor showing. The significant statistical test for differences among the subject's slope scores is complemented by the correlation between RT1 and RT4 of the letter-search task. Whereas the memory scanning results (Experiment 1) indicated that subjects' standard scores were quite similar in one and four-target memory scanning, the average correlation between RT1 and RT4 ($\bar{r} = .34$) indicates that subjects' standard scores were much less similar in one and four-target letter search. Even when the correlations are corrected for attenuation, the common one and four-target letter search true score variance appears to be only about one third of the total true score variance; the comparable figure was 96 percent common true score variance for memory scanning. High commonality between scores at extremes of the independent variable would indicate similar ordering of subjects at each extreme, so the slopes between the extremes would be similar for all subjects. The lack of commonality between one and four-target letter search scores indicates that subjects' slopes differed. Despite this, the reliability of the slope score was poor compared with the reliability of the RT scores for this letter-search task.

Table 11: Letter Search Intertrial Correlations (X 100) of Response Time: 1-Target Above Diagonal, 4-Target Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		38	80	43	67	57	71	66	63	63	81	71	71	63	76
2	61		30	37	28	52	33	51	22	38	43	24	61	49	61
3	73	74		32	78	43	50	48	49	47	65	51	48	57	50
4	57	65	69		24	55	41	56	46	47	40	39	55	19	61
5	65	62	73	55		27	44	40	26	26	54	52	35	66	35
6	39	43	61	72	43		63	75	63	65	64	61	75	41	66
7	67	26	65	54	54	58		78	74	66	78	80	77	65	78
8	41	09	20	34	49	07	40		75	76	86	78	78	62	78
9	85	60	75	62	70	44	54	36		67	74	71	68	38	67
10	19	26	37	44	41	11	22	38	25		75	53	66	43	77
11	54	36	69	46	72	51	56	51	61	35		80	74	69	72
12	52	32	36	32	62	18	26	56	56	43	63		67	80	65
13	30	-07	18	08	39	-09	45	77	21	22	47	28		60	90
14	36	17	15	21	45	-08	13	65	39	48	45	63	47		63
15	69	41	62	46	59	15	50	43	75	42	55	46	45	52	

Table 12: Letter Search Intertrial Correlations (X 100) of the Slope Score

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	11	34	04	42	20	60	41	37	22	25	47	49	36	51
2		21	42	27	15	-08	43	-15	-24	18	01	-29	03	06
3			05	37	-23	02	71	34	06	32	24	13	03	18
4				55	16	35	25	14	07	51	18	19	21	32
5					07	38	36	44	47	51	37	36	28	37
6						17	-00	-03	11	-05	25	07	03	-16
7							14	48	35	40	48	67	61	82
8								14	-10	59	29	16	17	34
9									52	33	48	59	37	50
10										36	70	61	49	34
11											47	59	60	47
12												62	66	45
13													76	65
14														60

Experiment 5: Search for Typographic Errors in Prose

Method

Reading is an iterative task which has not received much attention in the information processing literature. Schindler (1978) described a proofreading task for which he measured the frequency of failures to find typographic errors. He found a difference in results for typographic errors in function words (prepositions, conjunctions, auxiliary verbs, the verb "to be", pronouns, and possessive adjectives) versus content words (all others). Schindler found that function words were associated with significantly more failures to find the error, and that the effect was enhanced by a prose context.

In the present experiment, following Schindler, stimuli were extracted from Reader's Digest. Paragraphs of six equal lines were stored on photographic slides. There were 12 unrelated paragraphs for each of 15 trials of the experiment at daily intervals. Typographic errors were embedded, in the manner of Schindler (1978), in a content or a function word chosen at random on one of the six lines of each of the 12 paragraphs. Hence there were two types of misspelled words (content and function) representing each of six lines of prose on each day. The prose was in white letters on a blue background, and the lower case letters subtended .34 degrees at the subjects' sitting distance of 1m. from the projection screen.

In contrast to Schindler's measurement of error rates, time to find the typographic errors was measured in this experiment. Subjects were tested individually. The subject pressed a control button to display a paragraph on the screen (and start a timer). The subject was to release the button to extinguish the display (and stop the timer) when the typographic error was located. The subject then confirmed that he had found the error by indicating verbally what word was misspelled. The order of the lines on which the error appeared and the content-versus-function-word errors were randomized in the 12 slides of each of the 15 forms of the task.

There were seventeen subjects. Fifteen were in a Latin square design so that no two subjects saw the same form on the same day, and no subject saw the same form twice. The two remaining subjects saw the forms in random order.

Results

Means and standard deviations across all days and subjects are given in Table 13. The pattern shown in Table 13 was the same on all 15 days of the experiment ($F(70,1120) = .73$, $p = .95$). There was, however, an overall linear reduction of response times by about .2 seconds per day which accounts for 67 percent of the variance due to days ($F(1,16) = 33.03$, p .0005). None of the nonlinear trends across the 15 days of practice reached statistical significance (p .07). Typographic errors in content words took longer to detect than those in function words ($F(1,16) = 10.68$, $p = .005$). This effect was invariant across days ($F(14,224) = 1.18$, $p = .29$), but it changed depending upon the line of the paragraph in which the typographic error was located ($F(5,80) = 17.17$, p .0005). Function word misspellings were found faster in early lines, and content word errors were found relatively faster nearer the end of a paragraph. The steeper trend

for function words may be caused by a tendancy to overlook misspellings of function words (Schindler, 1978), thus necessitating rereading the paragraph and increasing the time required per line. Perhaps of greatest relevance to our discussion of slope scores is the fact that the linear trend (slope) of response time versus line number ($F(1,16) = 278.5$, p .0005) explained 96 percent of the variance attributable to line number.

Table 13: Means and Standard Deviations from a Typographic-Error Search Experiment (Experiment 5)

Word Type	Line Number					
	1	2	3	4	5	6
Content	6.2(9.0)	9.0(9.0)	10.7(8.6)	9.5(6.6)	11.0(6.3)	12.6(6.8)
Function	3.4(6.8)	5.0(4.7)	7.5(5.7)	8.2(4.5)	12.1(6.8)	13.4(7.6)

Note: Time to find errors, in seconds; standard deviations in parentheses. Data averaged across all days of a 15-day experiment.

Nonetheless, there were statistically significant third and fifth-order trends ($F(1,16) = 14.4$ and 29.6, respectively, p .0002). The effect of line number on response time ($F(70,1120) = .61$, $p = .995$) was invariant over days.

During initial consideration of individual differences in this task, it was noted that some response times were far too rapid or slow, compared with other response times for the same subjects on the same days. Median scores were adopted as a method for eliminating the outliers from consideration. The intertrial correlation matrix for median response time for function words is shown in the upper echelon of Table 14 ($\bar{r} = .37$), and the matrix for content words is in the upper echelon of Table 15 ($\bar{r} = .49$).

The outliers would have had too much influence in a calculation of least squares slopes, so a slope score that is less likely to be influenced by outliers was devised. It was based on the definition of slope: change in the dependent variable divided by change in the independent variable. With six levels of the independent variable there are 15 ways to calculate the slope: RT2-RT1, RT3-RT2, ..., (RT3-RT1)/2, (RT4-RT2)/2, ..., (RT4-RT1)/3, ..., (RT6-RT1)/5, where RTi is response time at level i of the independent variable. Some of these estimates will be too large or small due to outliers, but their median should accurately represent the slope. The intertrial correlation matrices of the median slope scores are shown in the lower echelons of Tables 14 and 15, respectively for function words ($\bar{r} = .21$) and content words ($\bar{r} = .03$). These correlations are generally weaker compared with the reliabilities of the median RT scores with the same number of degrees of freedom.

Table 14: Intertrial Correlations (X 100) for Function-Word Typographic-Error Search: Median Score Above Diagonal, Median Slope Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		26	47	10	21	-05	41	-13	20	23	58	09	58	42	15
2	51		40	22	49	76	43	-04	20	67	69	52	41	65	51
3	31	44		-07	50	15	34	-04	38	42	68	50	30	34	30
4	39	37	-18		18	07	47	-04	51	52	30	15	47	29	26
5	43	49	19	34		36	31	30	49	62	67	79	39	61	63
6	11	-17	32	-29	03		15	-17	08	38	41	26	10	30	19
7	-14	15	14	09	-23	-55		-20	52	59	57	23	31	39	37
8	36	00	-04	29	27	-35	17		37	-03	-02	23	10	10	35
9	11	37	04	15	56	-24	28	26		43	56	20	23	31	20
10	21	45	22	16	49	-20	25	17	69		72	76	67	59	71
11	47	64	32	06	59	-16	05	16	68	50		62	55	79	47
12	47	50	03	44	33	-27	45	39	53	59	45		53	54	77
13	-03	29	34	-12	53	-05	-14	07	50	42	47	-06		49	46
14	30	64	22	11	34	06	00	-04	42	53	57	13	46		53
15	21	25	47	28	10	-44	-04	24	-02	35	00	26	23	17	

Table 15: Intertrial Correlations (X 100) for Content-Word Typographic-Error Search: Median Score Above Diagonal, Median Slope Below Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		55	29	21	36	73	14	56	30	70	60	57	65	25	40
2	-30		18	73	60	59	22	67	57	71	79	64	71	41	64
3	-10	-02		52	47	04	-05	40	-09	42	08	40	32	32	15
4	-09	02	-43		74	41	28	69	30	72	61	63	66	52	60
5	-57	00	23	-06		37	-03	82	21	75	67	69	64	51	37
6	30	-26	-06	24	14		37	64	30	75	70	60	79	51	72
7	37	-40	-20	22	-08	-07		15	17	35	36	28	36	35	30
8	29	00	-10	-30	-60	04	-26		33	90	79	94	86	46	51
9	00	-24	-05	-33	29	-17	02	01		26	56	36	32	11	24
10	-21	43	18	-13	03	13	-21	39	-45		80	90	92	52	62
11	05	58	23	21	09	21	-25	00	-01	40		73	86	45	47
12	00	55	-19	-06	-17	28	-41	35	-17	47	26		88	36	53
13	-06	16	-23	25	04	65	-31	17	-40	34	21	63		43	64
14	12	-40	38	-02	30	52	14	-20	-22	-01	09	-37	22		66
15	-16	25	50	-23	18	-15	-25	16	37	28	51	18	-07	-20	

Experiment 6: Choice Reaction Time

Method

Fifteen subjects performed choice reaction time using a device identical to that used by Teichner (1978). One, two, and four-choice reaction times (without movement time) were measured in that order in three blocks of fifty responses. Blocks were presented in random order across days. Reaction times were measured at 10 to 15-second intervals, and a few minutes of rest were taken between blocks. Data were gathered on 15 successive weekdays, at approximately the same time on each day for any particular subject. Autocorrelation evidence was consistent with the simple assumption that successive reaction times within a block were statistically independent, so the average response time in each block was used as the performance score. Least-squares regression slopes of average reaction time as a function of the number of bits of information in the choice were calculated for each subject on each day.

Results

There was a statistically significant effect of the number of bits in the choice, $F(2,28) = 88.44$, p .0005 (see Table 16). The effect was invariant over days, $F(28,392) = 1.04$, p .401. Although the quadratic trend component was also statistically significant ($F(1,14) = 11.64$, $p = .004$), 98.5 percent of the variance was attributable to the linear trend ($F,(1,14) = 98.48$, p .0005). Hence the slope score would appear to represent an appropriate model.

Table 16: Means and Standard Deviations from a Choice Reaction Time Experiment (Experiment 6)

Number of Bits in the Choice		
0	1	2
186 (35)	216 (25)	233 (25)

Note: Reaction time in milliseconds; standard deviations in parentheses. Data averaged over the last 8 days of a 15-day experiment (Krause, Bittner, & Carter, 1982).

Days of practice had a predominantly (72 percent of the variance) linear effect ($F(1,14) = 54.86$, p .0005), with second and third order trends also reaching statistical significance ($F(1,14)$, = 21.21, p .0005, and $F(1,14) = 11.40$, $p = .0005$, respectively). The linear improvement with practice was about 12 milliseconds per day.

A test of homogeniety of variance based on the statistical jackknife (Miller, 1968) was adapted to multifacet ANOVA data, and it indicated no evidence for inhomogeneity over choices ($F(2,28) = 2.36$, $p = .113$), days ($F,(14,196) = .37$, $p = .98$), or their interaction ($F,(28,392) = .85$, $p = .69$).

Intertrial correlations were calculated to represent the reliability of reaction times (response times) to one and four-choice stimuli. These correlations are displayed in Table 17, $\bar{r}$ = .59 for responses to one-choice stimuli (above diagonal), and $\bar{r}$ = .57 for four-choice responses (below diagonal). Intertrial correlations representing the reliability of the slope scores are shown in Table 18, $\bar{r}$ = .39. The average correlation between RT1 and RT4 was .46, or .60 correlated for attenuation. As in the letter-search task, this low correlation indicates that subjects' slopes are different. Again slope scores are less reliable than the response times of which they are composed. This is true even though the slopes are based on more data than the mean RT with which they are being compared.

Table 17: Intertrial Correlations (X 100) of Choice Reaction Time: One-Choice Above the Main Diagonal, Four-Choice Below the Diagonal

Days	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		75	50	82	78	67	73	62	45	42	54	39	38	54	36
2	59		70	77	71	62	63	55	34	34	38	24	49	66	34
3	55	78		68	58	48	62	66	54	52	53	23	47	76	53
4	58	58	88		74	69	80	82	67	71	45	48	42	66	48
5	36	45	68	74		87	81	73	58	53	48	37	61	68	34
6	31	27	77	80	73		63	68	41	45	42	33	58	51	25
7	18	9	56	65	61	84		92	76	71	64	46	58	76	63
8	27	26	69	71	53	85	85		82	81	58	57	56	72	54
9	50	46	69	72	53	70	61	70		96	62	69	55	70	49
10	23	42	73	72	57	70	68	75	80		63	73	62	68	54
11	21	15	52	60	58	78	74	78	62	64		55	71	71	68
12	21	5	43	49	48	72	74	76	61	54	88		61	65	52
13	31	22	43	52	44	57	61	66	71	61	88	87		75	60
14	10	6	42	49	50	70	70	70	65	62	90	90	93		81
15	19	10	33	45	38	52	58	67	68	55	81	82	89	85	

Table 18: Choice Reaction Time: Slope Reliabilities over 15 Days (n=15)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	41	01	23	29	04	06	02	31	-26	-02	-12	36	-01	-09
2		23	49	45	24	36	23	35	-12	-14	-17	23	08	03
3			34	14	12	-14	-28	-02	16	01	-12	26	15	09
4				46	30	26	-09	07	-13	02	-13	-01	-18	08
5					61	61	53	72	31	39	34	60	33	50
6						66	73	55	43	60	69	75	55	51
7							83	84	59	70	66	60	62	70
8								77	60	73	80	69	74	64
9									60	65	56	76	67	63
10										75	75	52	76	72
11											88	59	75	87
12												61	80	84
13													67	47
14														72

DISCUSSION

As indicated by the six experiments presented above, slope scores are consistently less reliable than the response times from which the slopes are calculated. Even though combining several RTs improves the reliability of an average score, the combination of RTs in a slope apparently reduces the reliability of the composite. Several reasons for the unreliability of the slope measure have been described in the introduction and exemplified throughout the paper. Consequently, it is suggested that the slope score for individuals does not reliably characterize human information processing.

If slope scores must be used, however, there are ways to improve the reliability. In the remaining portion of this paper, these improvement methods will be described, followed by a discussion of some implications of the unreliability of slope scores for experiments and measurement of individual differences. Finally, it is suggested that most common research questions should be answered by analyzing RTs, rather that slope scores calculated for each subject.

Possible Ways to Alleviate Unreliability of Individual's Slopes

Methods to improve the reliability of the slope scores are suggested by the possible causes of unreliability listed previously. Several of these causes are fundamental and characteristic of slopes; it should be impossible to completely mitigate their effects. Nevertheless, reliabilities of slopes may be improved as follows.

Unreliability due to incorrectness of the linear model may be minimized by using slope scores only after verifying that the extra sum of squares (Draper & Smith, 1966) for nonlinearities is not statistically significant. This procedure may be used to test for the linearity of the composite slope, including all subjects, or to test for the linearity of each subject's data if sufficient degrees of freedom are available.

Similarly, unreliability due to homogeniety of the subjects' slopes may be overcome by verifying that there are at least detectable differences among the subjects' slopes. A conservative statistical test of this is the F-ratio of the mean square for subjects' slopes divided by the mean square for the subject-by-trials interaction. However, statistical significance of individual differences is a weak requirement. To ensure that the differences in slopes are considerable, one needs an index of the portion of the reliable variance of the RT which is not common between the extremes of the independent variable.

Unreliability due to the sensitivity of slopes to outliers may be reduced by preventing, trimming, or overwhelming the outliers. Outliers may be prevented by careful experimentation, including adequate preparation of subjects so that they may be motivated and trained to perform consistently. Outliers may be trimmed by using robust point estimators, such as the median, which gives no weight to extreme scores. Some robust estimators of slopes are suggested by robust estimators of correlations given by Devlin et al. (1975). Finally, outliers may be overwhelmed by large samples of response times at each level of the independent variable. Assuming that the data are not biased, the mean RT can be made as precise as desired by invoking the inverse square-root relationship between the standard deviation of a mean and its number of degrees of freedom.

The relation between slope scores and difference scores suggests that Cronbach and Furby's (1970) recommendations for estimation of differences be applied to slopes. Due to the intrinsic unreliability of differences, their first recommendation is not to estimate them. If, however, one must employ differences (slopes), the covariance among the data at various levels of the independent variable should be taken into account. Hence, if one had data RT1, RT2, and RT3 at equally spaced levels 1, 2, and 3 of an independent variable, an alternative to calculating slope: (RT3-RT1)/2 (which is equivalent to the least squares solution), would be to represent the information processing rate by RT3 with RT1 as a covariate. More complicated uses of covariance analysis to improve the reliabilities of slope scores are suggested in Cronbach and Furby's (1970) article on difference scores. Of course it is always possible to improve the reliability of differences or slopes by improving the reliability of the component response times. If these are mean response times, their reliability will improve in response to increased sample size as described by the Spearman-Brown formula (Winer, 1971).

Implications for Experiments and Measurement of Individual Differences

Experiments

It has recently been argued that the reliability of the dependent variable is not directly relevant to the power of a statistical test (Nicewander & Price, 1978). While this may be true for independent groups experiments, the power of a repeated-measures experiment increases with the reliability of the dependent variable (Sutcliffe, 1980). Hence, at least for repeated-measures experiments, a more reliable dependent variable is to be preferred to a less reliable one that represents the same thing. With this in mind, consider the following proposition. It is not necessary to calculate individual subjects' slope scores in an experiment with the hypothesis that slopes are affected by some treatment. The treatments may be variations of display or equipment design (e.g., Schiflett, 1980), or a stressor such as chemical exposure (Smith & Langolf, 1981), alcoholism (e.g., Mohs, Tinklenberg, Roth & Kopell, 1980), aging (e.g., Anders, Fozard, & Lillyquist, 1972), or brain injury (e.g., Harris & Fleer, 1974). When analyzing such data it is sufficient to deal directly with the more reliable response time scores. The hypothesis that slopes are altered by the treatment is tested by the F ratio for the treatments-by-(linear) conditions interaction, where conditions refers to positive set size, line number, number of targets, number of bits in a choice, and the like. A similar recommendation is made by Cronbach and Furby (1970) with respect to difference scores; it is not necessary to calculate differences or slopes of individual subjects in order to test hypotheses about differences or slopes. In fact, dealing directly with the human information processing response times has two advantages. First, a more powerful repeated-measures experiment will result from consideration of the more reliable response times. Second, consideration of response times does not restrict one to the linear model assumed in the slope calculation. The F ratio for the treatments-by-conditions interaction will indicate any change in human information processing, whatever the shape of the relation is between response time and conditions.

Individual Differences

Sometimes a researcher's objective is to show correlations between dependent variables representing individual differences. For example, Ford, Roth, Mohs, Hopkins, and Kopell (1979) showed correlations between event-related brain potentials and performance on Sternberg's (1966) task. Similarly, Cavanaugh (1972) suggested a correlation across types of memory material between memory span and memory scanning rate. Smith (1979) found a correlation across individuals between memory span and memory scanning rate. In such cases, it may be better to use regression techniques on the response times than to estimate the slope for each subject. For example, suppose it is hypothesized that there is a relation between variable W and the rate of processing in a 3-level information processing task with response times RT1 and RT3 at extremes of the independent variable. First RT3 can be regressed on RT1. Then the extra sum of squares (Draper & Smith, 1966) for adding W to the equation represents the relation between W and processing rate. Appropriate statistical significance tests are easily constructed. Other regression-related techniques suggested by Cronbach and Furby (1970) for correlation research involving difference scores can be adapted for information processing response times.

Conclusions

In general, there appears to be no practical reason to estimate information-processing slope scores for individuals. Since slope scores for individuals are both unreliable and unnecessary for ordinary purposes, it is suggested that only rare circumstances will justify their use.

ACKNOWLEDGEMENTS

This work was accomplished under Naval Medical Research and Development Command work unit No. MF58.524-002-5027. The project was entitled Performance Evaluation Tests for Environmental Research (PETER). The authors wish to acknowledge the special contributions of Robert Kennedy who, acting as principal investigator of the work unit, directed that the experiments be conducted and suggested that a paper be written on the reliability of slope scores. We also wish to acknowledge the many people who helped to gather the data: Susan Jones, Michael Shewmake, Mary Harbeson, Debra Andrews, Yvonne Boudreau, and Lawrence Bell. Andrew Rose, Harold Hawkins, and Robert Sekuler gave advice and assistance at various stages of this project. Richard Shannon and Alvah Bittner, Jr. made helpful suggestions for improvement of the manuscript.

References

Anders, T. R., Fozard, J. L., & Lillyquist, T. D. Effects of age upon retrieval from short-term memory. Developmental Psychology, 1972, 6, 214-217.

Burrows, D., & Murdock, B. B. Effects of extended practice on high-speed scanning. Journal of Experimental Psychology, 1969, 82, 231-237.

Cavanagh, J. P. Relation between the immediate memory span and the memory search rate. Psychological Review, 1972, 79, 525-530.

Collins, A. M., & Quillian, M. R. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 240-247.

Cronbach, L. J., & Furby, L. How we should measure change-or should we? Psychological Bulletin, 1970, 74, 68-80.

David, H. A. Upper 5 and 1% points of the maximum F-ratio. Biometrika, 1952, 39, 422-424.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. Robust estimation and outlier detection with correlation coefficients. Biometrika, 1975, 62, 531-545.

Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Fernandes, K., & Rose, A. M. An information processing approach to performance assessment: II. An investigation of encoding and retrieval processing in memory. (Tech. Report AIR 58500-11/78 TR). Washington, D. C.: American Institutes for Research, November, 1978.

Ford, J. M., Roth, W. T., Mohs, R. C., Hopkins, W. F., III, & Kopell, B. S. Event-related potentials recorded from young and old adults during a memory retrieval task. Electroencephalography and Clinical Neurophysiology, 1979, 47, 450-459.

Harris, G., & Fleer, R. High speed memory scanning in mental retardates: Evidence for a central processing deficit. Journal of Experimental Child Psychology, 1974, 17, 452-459.

Krause, M., Bittner, A. C., Jr., & Carter, R. C. Repeated measures on a choice reaction time task, 1981, manuscript in preparation.

Kristofferson, M. W. When item recognition and visual search functions are similar. Perception and Psychophysics, 1972, 12, 379-384.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley publishing Co., 1968.

Miller, R. G., Jr. Jackknifing variances. Annals of Mathematical Statistics, 1968, 39, 567-582.

Mohs, R. C., Tinklenberg, J. R., Roth, W. T., & Kopell, B. S. Slowing of short-term memory scanning in alcoholics. Journal of Studies on Alcohol, 1978, 39, 1908-1915.

Neisser, U. Decision time without reaction time: experiments in visual scanning. American Journal of Psychology, 1963, 76, 376-385.

Neisser, U., Novick, R., & Lazar, R. Searching for ten targets simultaneously. Perceptual and Motor Skills, 1963, 17, 955-961.

Nicewander, W. A., & Price, J. M. Dependent variable reliability and the power of significance tests. Psychological Bulletin, 1978, 85, 405-409.

Posner, M. I. Chronometric explorations of mind. New York: Wiley, 1978.

Ross, J. Extended practice with a single character classification task. Perception and Psychophysics, 1970, 8, 276-278.

Schiflett, S. G. Evaluation of a pilot workload assessment device to test alternate display formats and control handling qualities (SY-33R-80). Patuxent River, MD: Naval Air Test Center, July 1980.

Schindler, R. M. The effect of prose context on visual search for letters. Memory & Cognition, 1978, 6, 124-130.

Simpson, P. J. High speed scanning: Stability and generality. Journal of Experimental Psychology, 1972, 96, 239-246.

Smith, P. J. Short-term memory scanning is related to memory scan and mercury exposure. PhD. dissertation, The University of Michigan, 1979.

Smith, P. J., & Langolf, G. D. The use of Sternberg's memory scanning paradigm in assessing effects of chemical exposure. Human Factors, 1981, 23, 701-708.

Sternberg, S. High speed scanning in human memory. Science, 1966, 153, 652-654.

Sternberg, S. Memory scanning: Mental processes revealed by reaction-time experiments. American Scientist, 1969, 57, 421-457.

Sternberg, S. Memory scanning: New findings and current controversies. Quarterly Journal of Experimental Psychology, 1975, 27, 1-32.

Sutcliffe, J. P. On the relationship of reliability to statistical power. Psychological Bulletin, 1980, 88, 509-515.

Teichner, W. H., Personal Communication, 1978.

Thomas, D. J., Majewski, P. L., Ewing, C. L., & Gilbert, N. S. Medical qualification procedures for hazardous-duty aeromedical research. (Conference Proceedings No. 231) Nuilly-Sur-Seine, France: AGARD,

END

FILMED

8-83

DTIC